Review

Literature review of data mining applications in academic libraries

Authors:

Lorena Siguenza-Guzman[a,b,*], Victor Saquicela[a], Elina Avila Ordóñez[a,b], Joos Vandewalle[c], Dirk Cattrysse[b]

Affiliations:

[a] Department of Computer Science, University of Cuenca. 12 de Abril Av., ECU-010150, Cuenca – Ecuador
[b] Centre for Industrial Management Traffic & Infrastructure, KU Leuven. Celestijnenlaan 300, Box 2422, BE-3001, Leuven – Belgium
[c] Department of Electrical Engineering ESAT/Stadius, KU Leuven. Kasteelpark Arenberg 10, Box 2440, BE-3001, Leuven – Belgium

Email addresses:

lorena.siguenza@ucuenca.edu.ec, victor.saquicela@ucuenca.edu.ec, elina.avilao@ucuenca.edu.ec, joos.vandewalle@kuleuven.be, dirk.cattrysse@kuleuven.be

Corresponding author:

* Celestijnenlaan 300 - Box 2422. Office: 04.44| BE-3001 Leuven | Belgium

Tel.: + 32 (0)16 37 27 65 | Fax: + 32 (0)16 32 29 86 | Mob (GSM): + 32 (0) 484 26 50 03

Alternative email: lorena.siguenzaguzman@kuleuven.be | Website: http://cib.kuleuven.be

Literature review of data mining applications in academic libraries

Abstract

This article provides a comprehensive literature review and classification method for data mining techniques applied to academic libraries. To achieve this, forty-one practical contributions over the period 1998-2014 were identified and reviewed for their direct relevance. Each article was categorized according to the main data mining functions: clustering, association, classification, and regression; and their application in the four main library aspects: services, quality, collection, and usage behavior. Findings indicate that both collection and usage behavior analyses have received most of the research attention, especially related to collection development and usability of websites and online services respectively. Furthermore, classification and regression models are the two most commonly used data mining functions applied in library settings.

Additionally, results indicate that the top 6 journals of articles published on the application of data mining techniques in academic libraries are: College and Research Libraries, Journal of Academic Librarianship, Information Processing and Management, Library Hi Tech, International Journal of Knowledge, Culture and Change Management, and The Electronic Library. Scopus is the multidisciplinary database that provides the best coverage of journal articles identified. To our knowledge, this study represents the first systematic, identifiable and comprehensive academic literature review of data mining techniques applied to academic libraries.

Keywords

Data mining, bibliomining, literature review, classification, clustering, association, regression, academic libraries

Literature review of data mining applications in academic libraries

1. Introduction

Data mining, also known as knowledge discovery in databases, can be defined as the process of analyzing large information repositories and of discovering implicit, but potentially useful information (Han, Kamber, & Pei, 2011). Data mining has the capability to uncover hidden relationships and to reveal unknown patterns and trends by digging into large amounts of data (Sumathi & Sivanandam, 2006). The functions, or models, of data mining can be categorized according to the task performed: association, classification, clustering, and regression (Hui & Jha, 2000; Kao, Chang, & Lin, 2003; Nicholson, 2006b).

Data mining analysis is based normally on three techniques: classical statistics, artificial intelligence, and machine learning (Girija & Srivatsa, 2006). *Classical statistics* is mainly used for studying data, data relationships, as well as for dealing with numeric data in large databases (David J. Hand, 1998). Examples of classical statistics include regression analysis, cluster analysis, and discriminate analysis. *Artificial intelligence* (AI) applies "human-thought-like" processing to statistical problems (Girija & Srivatsa, 2006). AI uses several techniques such as genetic algorithms, fuzzy logic, and neural computing. Finally, *machine learning* is the combination of advanced statistical methods and AI heuristics, used for data analysis and knowledge discovery (Kononenko & Kukar, 2007). Machine learning uses several classes of techniques: neural networks, symbolic learning, genetic algorithms, and swarm optimization. Data mining benefits from these technologies, but differs from the objective pursued: extracting patterns, describing trends, and predicting behavior.

A typical data mining process, as shown in Figure 1, is an interactive sequence of steps that normally starts by integrating raw data from different data sources and formats. These raw data are cleansed in order to remove noise, and duplicated and inconsistent data (Han et al., 2011). These cleansed data are then transformed into appropriated formats that can be understood by other data mining tools, and filtration and aggregation techniques are applied to the data in order to extract summarized data. In fact,

interesting knowledge is extracted from the transformed data. This information is analyzed in order to identify the truly interesting patterns. Eventually, knowledge is visualized to the user. More detailed information regarding a data mining process can be found in Han et al. (2011).

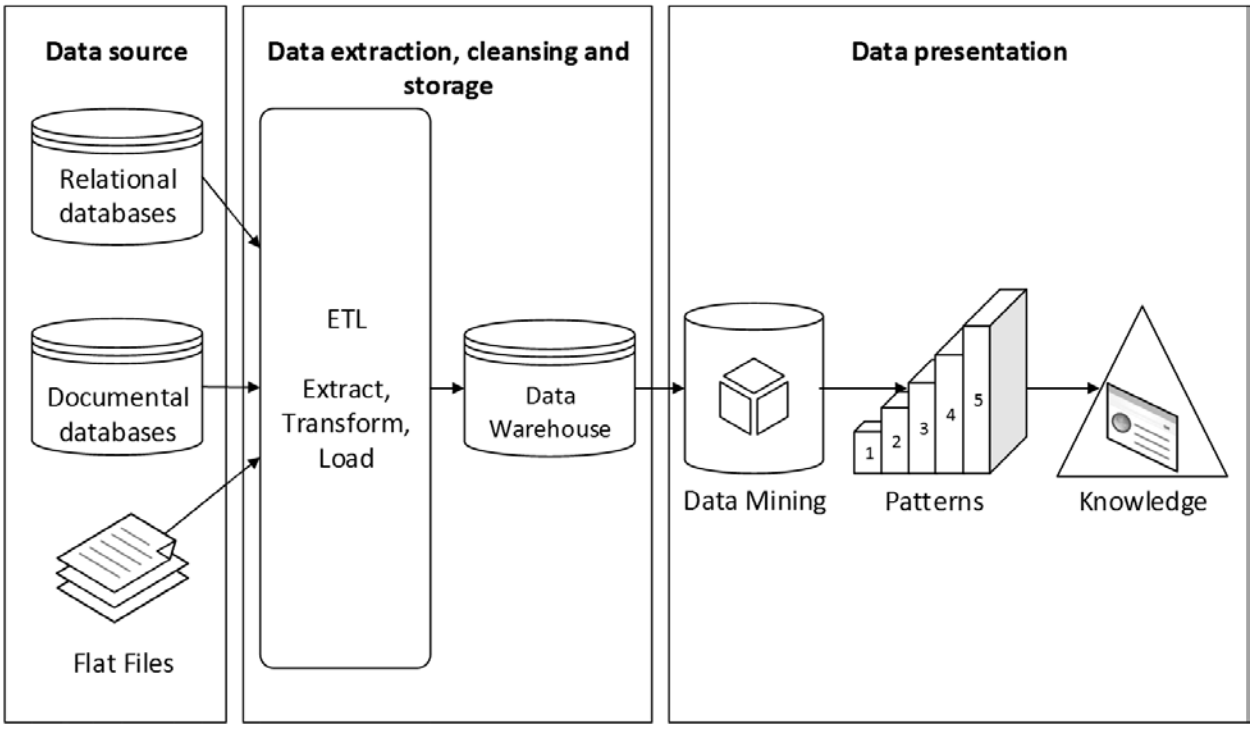

Figure 1: Data mining process, based on Han et al. (2011).

Data mining techniques are applied in a wide range of domains where large amounts of data are available for the identification of unknown or hidden information. In this sense, Girija and Srivatsa (2006) indicate that data mining techniques used in WWW are called web mining, used in text are called text mining, and used in libraries are called bibliomining.

The term bibliomining, or data mining for libraries, was first used by Nicholson and Stanton (2003) to describe the combination of data warehousing, data mining and bibliometrics. This term is used to track patterns, behavior changes, and trends of library systems transactions. Although the concept is not new, the term bibliomining was created to facilitate the search of the terms “library” and “data mining” in the context of libraries rather than in software libraries. Bibliomining is an important tool to discover useful

library information in historical data to support decision-making (Kao et al., 2003). However, to provide a complete report of the library system, bibliomining needs to be used iteratively applied in combination with other measurement and evaluation methods; as strategic information is discovered, more questions may be raised and thus start the process again (Nicholson, 2003b).

Bibliomining, as any knowledge extraction method, needs to follow a systematic procedure in order to allow an appropriate knowledge discovery. The bibliomining process starts by determining areas of focus and collecting data from internal and external sources (Nicholson, 2003b). Then, these data are collected, cleansed, and anonymized into a data warehouse. To discover meaningful patterns in the collected data, the bibliomining process includes the selection of appropriate analysis tools and techniques from statistics, data mining, and bibliometrics (Nicholson, 2006a). Interesting patterns are analyzed and visualized through reports. The mining process will be iterated until the resulted information is verified and proved by key users such as librarians and library managers (Shieh, 2010).

The application of bibliomining tools is an emerging trend that can be used to understand patterns of behavior among library users and staff, and patterns of information resource use throughout the library (Nicholson & Stanton, 2006). Bibliomining is highly recommended to provide useful and necessary information for library management requirements, focusing on the professional librarianship issues, but highly database technical dependent (Shieh, 2010). Bibliomining can also be used to provide a comprehensive overview of the library workflow in order to monitor staff performance, determine areas of deficiency, and predict future user requirements (Prakash, Chand, & Gohel, 2004). The resulting information gives the possibility to perform scenario analysis of the library system, where different situations that need to be taken into account during a decision-making process are evaluated (Nicholson, 2006a). An additional application is to standardize structures and reports in order to share data warehouses among groups of libraries, allowing libraries to benchmark their information (Nicholson, 2006a). Therefore, in order to improve the interaction quality between a library and its users, the application of data mining tools in libraries is worth pursuing (Chang & Chen, 2006).

The aim of this study is to investigate how far academic libraries are pragmatically using data mining tools, and in which library aspects librarians are implementing them. To this end, content and statistical analyses are used to examine articles that include case studies of academic libraries implementing data mining tools. The remainder of the article provides a detailed explanation of the research methodology adopted in this literature study. This is followed by a description of the proposed method for classifying data mining applications in libraries. Classification results are then presented and discussed. The article concludes by presenting limitations of the study, and by outlining research implications and prospects for future research.

2. Research methodology

The present study follows the methodology employed by Ngai et al. (2009) to analyze and classify data mining techniques applied to customer relationship management. In this study, the analysis and classification are based on the examination of selected search engines and the use of a set of descriptors, all related to their specific interests. Then, the selected articles are reviewed and categorized based on a classification framework. The resulting list and classification is independently verified by research triangulation; finally, findings are reported in order to identify implications and future research directions.

Thus, following the Ngai et al. selection criteria and evaluation framework, a Web-based literature research on practical documents about data mining applications was conducted in order to identify relevant articles. As the nature of research on data mining and libraries is difficult to comprehend within the confines of specific disciplines, the relevant articles are scattered throughout numerous scholarly journals. Consequently, bearing in mind the degree of relevance or specialization to the subject analyzed, a set of four search engines was first selected to perform journal browsing. Based on the specialization degree, two major Library and Information Science (LIS) databases were searched: Library Information Science & Technology Abstracts (LISTA) accessed through EBSCOhost, and Library and Information Science Abstracts (LISA) accessed through ProQuest. In addition, two multidisciplinary databases: Web

of Science (WoS) and Scopus were also consulted as complementary databases, as both search engines are among the largest and most common of the multidisciplinary databases available. Subsequently, citation tracing was also employed to discover additional papers relevant to this study; thus, the reference section for each article found was traced in order to find additional journal articles.

The search was operated according to the following procedure. First, a selection of subject terms was performed in order to identify terms that represent the concepts related to the topic under the study. To this end, the thesauruses of LISA and LISTA were consulted to draw up a set of standardized descriptors. Although the term “bibliomining” does not appear in both thesauruses, it was also incorporated as a subject term in order to investigate if academics and practitioners utilize this word as part of their titles or provided keywords. Based on this terms selection, relevant articles were searched by combining the following subject terms: “data mining”, “academic librar*” and “university librar*”. The asterisk (*) is used to find words ending with a common stem, for example, librar* = librar*ies* or library. All these search terms and their combinations were searched in subject headings (article title, abstract and keywords), and the analysis was limited to journal articles published in English. An overview of the criteria and results is shown in Table 1. When the number of articles searched for was within a reasonable number to conduct analysis, the resultant literature was sorted, summarized, and discussed in order to generate a final sample consisting of 485 potentially relevant studies. Then, the full text of each article was retrieved for detailed evaluation in order to eliminate those articles that did not meet the selection criteria with the application of data mining techniques in academic libraries. Each excluded article was registered in an excluded-studies table, followed by an explanation for its separation. All excluded articles were further screened by a different reviewer to confirm agreement with exclusion. In addition, the reference section for each included article was examined for possible titles of additional studies. By so doing, a total of 135 extra articles were analyzed. The standardized inclusion/exclusion criteria were as follows:

- Only English articles were included in the study.

- Only the articles related to the application of data mining techniques in academic libraries were selected, as these were the focus of this literature review.
- The articles describing the application of data mining techniques in academic libraries without a specific case study were excluded.
- Only the articles clearly describing how the mentioned data mining technique(s) could be applied and assisted in library settings were selected.
- Masters and doctoral dissertations, conference papers, text books and unpublished working papers were excluded. The main reason for this decision was that both academics and practitioners most commonly use journals both to acquire information and spread new knowledge (Gonzalez, Llopis, & Gasco, 2013). Whereas journal articles currently represent the highest level of research, other formats, like books, are confined to gathering and spreading knowledge that is already established. As for conferences, it is usual for most valuable articles to end up being published in journals; in fact, the conference represents a step prior to the definitive journal publication (Gonzalez et al., 2013).

Table 1

Search criteria and results number per database

Database	Query	Number of results
LISTA	bibliomining	11
	"data mining" "academic librar*"	20
	"data mining" "university librar*"	16
LISA	bibliomining	9
	"data mining" "academic librar*"	9
	"data mining" "university librar*"	11
Web of Science	bibliomining	14
	"data mining" "academic librar*"	45
	"data mining" "university librar*"	12
Scopus	bibliomining	31
	"data mining" "academic librar*"	140
	"data mining" "academic librar*" "case stud*"	53
	"data mining" "university librar*"	92
	"data mining" "university librar*" "case stud*"	22
	Total articles analyzed	*485*

Forty-one articles were subsequently selected. A detailed table of the selected articles can be found in Appendix A. Each selected article was carefully reviewed and separately classified according to four quadrants of a holistic evaluation matrix for libraries and four main data mining functions, as shown in Figure 2. Although this research was not exhaustive and oriented to the application of data mining techniques in academic/university libraries, it serves as a comprehensive base for an understanding of data mining research in libraries in general.

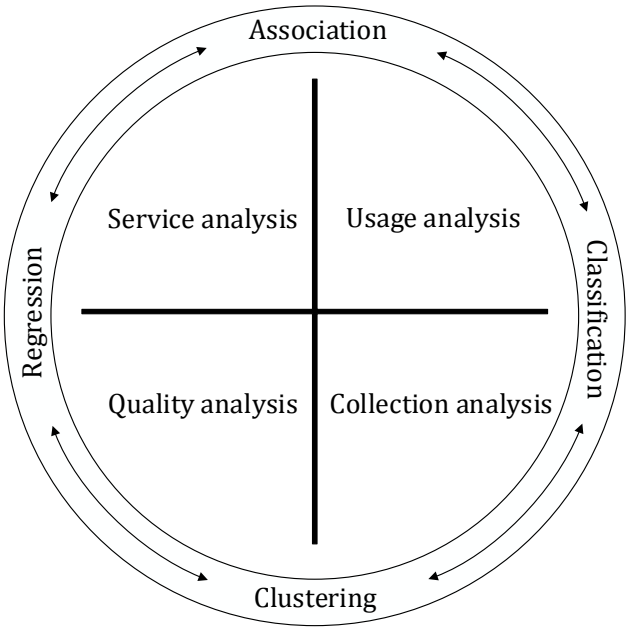

Figure 2: Classification framework for data mining techniques based on the Ngai et al. (2009) approach

3. Classification method

3.1 Classification framework: Holistic approach for library evaluation

Facing rapidly changing landscapes, characterized by shrinking budgets and dynamic services, libraries have recognized the need for evidence of their value. Academic libraries more than ever before, are called upon to demonstrate and justify their existence, and their contribution to institutional missions and goals (Association of College and Research Libraries, 2010). In fact, new trends and issues affecting academic libraries include a culture of increasing accountability for outcomes, in which libraries will be required to find better ways to document these connections (ACRL Research Planning and Review Committee,

2014). Nicholson (2004) and Siguenza-Guzman et al. (2015) recognizing the need to evaluate libraries in a holistic and structured manner, propose the use of a two-dimensional evaluation matrix. The four quadrants of the holistic evaluation matrix are the following:

1. *Internal perspective of the library system – Process/service analysis:* In this quadrant, the "library system" refers to everything that is part of the offerings of the library such as the organizational scheme, electronic equipment, library staff, and facilities. Internal perspective of the library system involves analyzing the topics related to processes and services carried out within the library.
2. *External perspective of the library system – Quality analysis:* Quality of the collection and services are assessed by users. Thus, the second quadrant evaluates the aboutness, pertinence, and usability of physical and digital resources by exploring users' perceptions (Nicholson, 2004). Assessment methods to measure the quality of services and collection include statistics gathering, suggestion boxes, Web usability testing, user interface usability, and satisfaction surveys (Wright & White, 2007).
3. *Internal perspective of the library collection – Collection analysis:* The third quadrant aims to evaluate the usefulness of the library collection. Proponents of this holistic approach suggest the combination of three assessment methods; namely, citation analysis, vendor-supplied statistics, and citation databases. By doing so, libraries will gain an extensive knowledge *about* their collection value and information relevance.
4. *External perspective of the library collection – Usage analysis:* This final quadrant evaluates users' behavior when manipulating the library system. Users' interaction with the system is utilized to study users' preferences to personalize library services. Transaction log analysis, Web usage analysis, deep log analysis, and usage statistics are the main techniques utilized for this purpose.

Each quadrant of this evaluation matrix shares the common goal of supporting libraries in gaining a thorough and holistic understanding of their users and services. Data mining techniques, therefore, can

help to accomplish such a goal by uncovering hidden patterns of behavior among library users and staff members, and patterns of information resource usage (Nicholson & Stanton, 2003).

3.2 Classification framework: Data mining models

Bibliomining can reveal issues associated with information-seeking user behavior, predict future trends on collection development, and build user communities based on common information interests. Based on the type of knowledge discovery, data mining functions can be divided into unsupervised and supervised algorithms (Chen & Liu, 2004). The former recognizes relationships in non-classified data, while the latter requires the data to be pre-classified in order to explain those relationships. According to these two main function types, data mining algorithms can be divided into the following categories: association, clustering, classification, and regression (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; D. J. Hand, Mannila, & Smyth, 2001).

1. *Association:* The so-called association rule aims to find the existing (or potential) relationships between data items in a database such as attributes and variables (Lunfeng, Huan, & Li, 2012). Examples of common association tools are statistics and apriori algorithms (Ngai et al., 2009).
2. *Clustering:* Clustering is the task of uncovering unanticipated trends by segmenting no predefined clusters. This approach is used in situations where a training set of pre-classified records is unavailable (Chen & Liu, 2004). Common tools for clustering include neural networks, k-means algorithms, and discrimination analysis (Ngai et al., 2009).
3. *Classification:* Classification is the task of attempting to discover predictive patterns by classifying database records into a number of predefined categorical classes based on certain criteria (Chen & Liu, 2004). Common classification tools are neural networks, decision trees, and if- then-else rules (Ngai et al., 2009).
4. *Regression:* Regression is an essentially statistical technique that maps a data item to a real-valued prediction variable. This data mining function is normally used to capture the trends of frequent

patterns. Examples of common regression techniques include linear regression and logistic regression analysis.

In turn, numerous data mining techniques are available for each type of data mining function. The choice of the data mining technique depends on the nature and purpose for the research study or the library requirements (Banerjee, 1998). Examples of some widely used data mining algorithms include the following: k-means algorithms for clustering, association rules for association, linear and logistic regression for regression, and decision trees for classification.

3.3 Research classification process

Following the Ngai et al. (2009) classification process approach, each selected article was reviewed and classified according to the proposed classification framework by three independent researchers. Researcher A was selected based on their expertise on the library holistic approach, whereas the other two B and C were selected based on their data mining experience. The classification process consists of five phases:

1) Online database search.
2) First classification by researcher B.
3) First verification of classification results and excluded articles by researcher A.
4) Independent verification of classification results by researcher C.
5) Discussion on classification results by the researchers A, B and C, and
6) Analysis and tabulation of results by the researchers A and B.

If a discrepancy in classification results existed between researchers, each article was then discussed until an agreement was reached by consensus on how the article should be classified from the final set in the proposed classification framework. Figure 3 shows the selection process utilized across the study. The collection of articles was analyzed based on the library holistic matrix and data mining models, by year of

publication, country of implementation, journal in which the article was published, as well as the type of library analyzed (physical, digital or both).

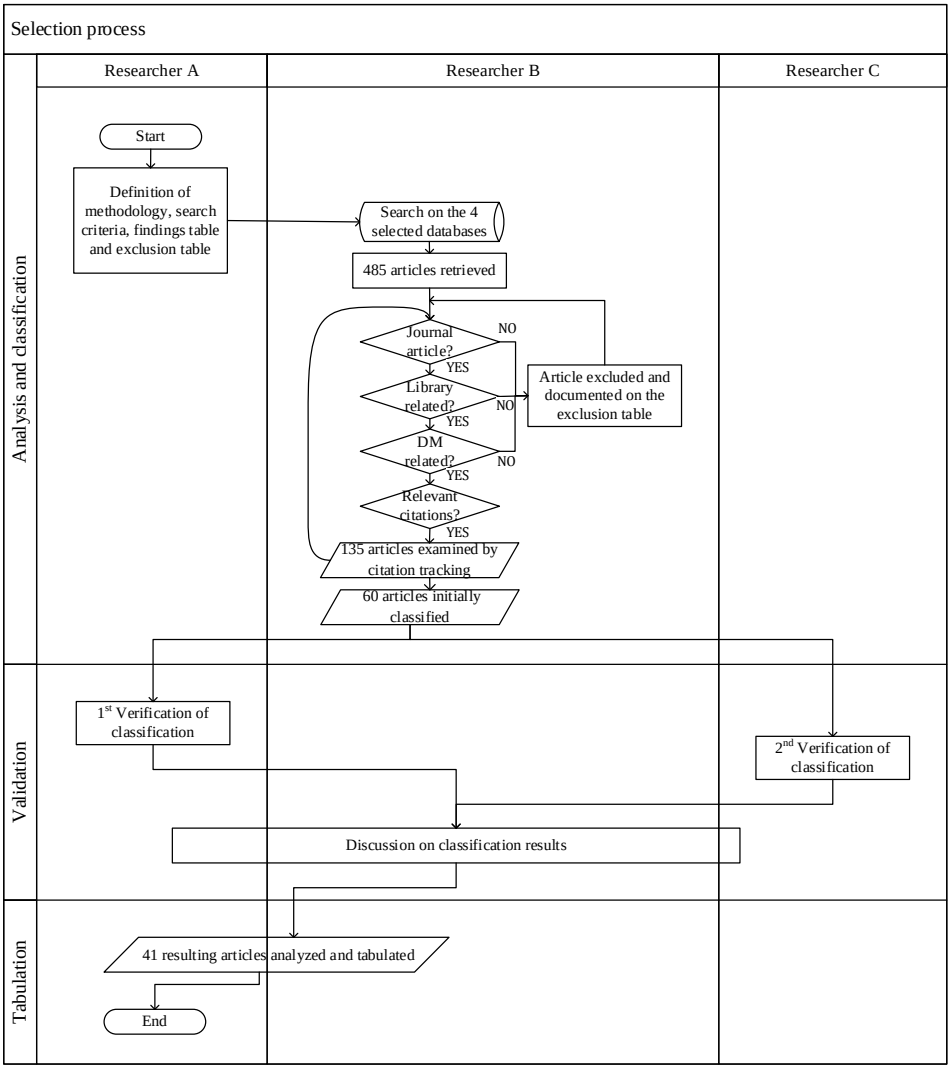

Figure 3: Selection process framework

4. Classification of the articles

A detailed distribution of the 41 articles classified by means of the proposed framework is shown in Table 2.

Table 2

Distribution of articles according to the proposed classification model

Library holistic evaluation	Data mining functions	Data mining techniques	References
Service analysis	Association	Association rules	Decker and Höppner (2006)
	Classification	Logical analysis of data	Leonard et al. (2010), Zweibel and Lane (2012), Sewell (2013)

	Clustering	k-means algorithm	Tempelman-Kluit and Pearce (2014)
	Regression	Linear regression	William (2007), Weiner (2009), Emmons and Wilkinson (2011)
		Logistic regression	Yi (2009), Yi (2011), Yi (2012)
		Regression analysis	Siriprasoetsin et al. (2011)
Quality evaluation	Classification	Neural network	Decker and Hermelbracht (2006), Papavlasopoulos and Poulos (2012)
	Regression	Linear regression	Whitmire (2002)
		Regression analysis	Siriprasoetsin et al. (2011)
Collection analysis	Association	Association rules	Wu et al. (2004), Zhang and Wang (2013), Li (2014)
		Bibliometric analysis	Will (2006)
		Statistical analysis	Tosaka and Weng (2011)
	Classification	Decision/Classification tree	Kao et al. (2003), Nicholson (2003a), Wu (2003), King et al. (2007), Yang (2012)
		Log analysis	Nicholas et al. (2006)
		Logical analysis of data	Leonard et al. (2010), Zweibel and Lane (2012)
		Logistic regression	Nicholson (2003a)
		Memory-based reasoning	Nicholson (2003a)
		Neural network	Nicholson (2003a), Papavlasopoulos and Poulos (2012)
	Clustering	Decision/Classification tree	Koulouris and Kapidakis (2012)
		Pattern based clustering	Shreeves et al. (2003)
	Regression	Linear regression	William (2007), Emmons and Wilkinson (2011)
		Logistic regression	Soria et al. (2014)
Usage analysis	Association	Association rules	Pu and Yang (2003), Wu et al. (2004), Decker and Höppner (2006), Zhang and Wang (2013)
		Log analysis	Blecic et al. (1998)
		Statistical analysis	Blecic et al. (1998), Tosaka and Weng (2011)
	Classification	Decision/Classification tree	King et al. (2007)
		Log analysis	Nicholas et al. (2006), Shieh (2012), Ahmad et al. (2014)
		Logical analysis of data	Samson (2014)
		Neural network	Decker and Hermelbracht (2006)
		Statistical analysis	Shieh (2012)
	Clustering	Hierarchical cluster analysis	Bollen and Luce (2002), Hájek and Stejskal (2014)
		k-means algorithm	Bollen and Luce (2002), Hájek and Stejskal (2014), Tempelman-Kluit and Pearce (2014)
		Logical analysis of data	Finnell and Fontane (2010)
		Pattern based clustering	Papatheodorou et al. (2003), Shreeves et al. (2003), Todorinova et al. (2011)
	Regression	Linear regression	Weiner (2009), Emmons and Wilkinson (2011), Fagan (2014)
		Logistic regression	Bracke (2004), Soria et al. (2014)

* Remark: Each article may have used more than one data mining technique and may have been implemented in more than one library holistic quadrant; thus, it may appear more than once

4.1 Distribution of articles by the library holistic quadrants and data mining models

The distribution of articles classified by the proposed classification model is shown in Table 3. It is striking that a large part of the published case studies on the use of data mining in libraries are case studies of usage behavior analysis (24 out of 41 articles, 59%). Of these 24 articles, almost 38% of the articles (nine in total) are related to the analysis or characterization of data. For instance, *log analysis* is reported in four articles to analyze the information seeking behavior in regard to digital libraries and library websites. Specifically, Blecic et al. (1998) employ transaction log analysis of an OPAC and statistical tools to improve information retrieval. Nicholas et al. (2006) report a deep log investigation of the use and users of the Blackwell Synergy, a proprietary interdisciplinary digital library. Ahmad et al.

(2014) utilize deep log analysis techniques to evaluate the user acceptance of e-book adoption. Shieh (2012) utilizes log analysis and statistical tools to evaluate the usability and findability of library websites. *Statistics tools* are employed in the total of three articles, which include two of the above-described studies: Blecic et al. (1998), Shieh (2012), and the study presented by Tosaka and Weng (2011) using statistical tools to examine the effect of content-enriched records on library materials usage. *Logical analysis of data* is discussed in two articles: Finnell and Fontane (2010), which employ these tools to investigate the feasibility of using reference questions as a tool in the construction of study guides, instructional outreach, and collection development. In a recent study, Samson (2014) analyzes the value of library resources to institutional teaching and research needs through the usage study of library e-resources.

More sophisticated data mining techniques used in this quadrant include: association rules (four out of 24 articles), linear regression (three articles in total), k-means algorithm (3 articles), and pattern based clustering (three articles). Regarding *association rules*, Pu and Yang (2003) provide new basis for information organization and retrieval applications. Authors utilize circulation patterns of similar users to discover association classes scattered across different subject hierarchies. Wu et al. (2004) use circulation statistics and association rule discovery to support decision-making for material acquisitions. Specifically, association rules are employed to open up the relationship between pairs of material categories to predict the users' needs. Decker and Höppner (2006) apply an association rules-based approach to explore the use of customer intelligence to support strategic planning processes using data warehouse tools. Zhang and Wang (2013) report the implementation of association rules to mine transactional data generated in the process of library service. The aim of this study is to provide accurate service for readers based on a user behavior analysis. *Linear regression models* are utilized by several authors to demonstrate library's value. For instance, Weiner(2009) utilize multiple regression analysis to analyze the library contribution to the University reputation, while Emmons and Wilkinson (2011) apply a linear regression model to evaluate the impact of academic libraries on students persistence. Fagan (2014) has recently used linear regression analyses to explore relationships among several variables thought to predict full-text article requests, such

as reference transactions, library instruction, database searches, and ongoing expenditures. Concerning the implementation of *k-means algorithms*, Bollen and Luce (2002) and Hájek and Stejskal (2014) report the implementation of two types of cluster analysis: hierarchical cluster analysis and k-means clustering to analyze user retrieval patterns in digital libraries. Bollen and Luce (2002) analyze the retrieval habits of users in order to assess the impact of a library collection and to determine the structure of a given user community. Hájek and Stejskal (2014) try to identify the user behavior of a typical consumer to support library management ensuring the provision of the appropriate level of library services. Tempelman-Kluit and Pearce (2014) utilize a k-means cluster to mine a Library 2.0 service. Chat reference data are analyzed to create hypothetical users (personas) that represent behaviors, goals and values of actual users. Eventually, Papatheodorou et al. (2003) use pattern based clustering to construct user communities sharing common interests and preferences. To do so, Z39.50 session log files are recorded and mined. Shreeves et al. (2003) identifies document clusters of potential interest, and provides visual displays of these clusters and document similarities. This study is part of a bigger project to examine the efficacy of using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to construct a search and discovery service focused on information resources in the domain of cultural heritage. Todorinova et al. (2011) examine the staffing patterns at the reference desk in order to give librarians greater flexibility as well as to allow better responding to the information-seeking needs of users.

Furthermore, 19 out of 41 articles (46%) deal with the application of data mining models in collection analysis, 12 articles (29%) with service analysis, and four articles (10%) with quality analysis, thus covering various aspects of library services and collection.

Articles covering the holistic quadrants of services, collection, and usage analysis apply all four data mining functions, whereas collection analyses do not employ cluster algorithms for their analyses. Collection and usage analyses are the two quadrants that have been the most explored together (nine articles). The majority of articles regarding quality analysis also cover the other three library aspects (three out of four articles). In fact, quality is the quadrant with the least-independent works, whereas usage behavior is the quadrant with the highest number of independent works (12 out of 24 articles).

Table 3

Distribution of articles by the library holistic quadrant and data mining models

Holistic evaluation quadrants	Number per holistic quadrant	Data mining functions	Amount per data mining function
Service analysis	12	Association	1
		Classification	3
		Clustering	1
		Regression	7
Quality evaluation	4	Classification	2
		Regression	2
Collection analysis	19	Association	5
		Classification	9
		Clustering	2
		Regression	3
Usage analysis	24	Association	6
		Classification	6
		Clustering	7
		Prediction	5

* Remark: Each article may have used more than one data mining technique and may have been implemented in more than one library holistic quadrant

Within the 24 articles of usage analysis, implementation of the data mining functions are almost equally distributed among them,; that is, seven articles (29%) use clustering models to analyze the usage behavior of library collection, followed by association models and classification rules that are both discussed in six articles (25%) each, and five articles (21%) which use regression models. Regarding to collection analysis, 47% (nine out of 19 articles) use classification models, and 26% (five articles) utilize association models. Figure 4 shows a visual representation of the classification of data mining applications based on the quadrants of the holistic evaluation matrix.

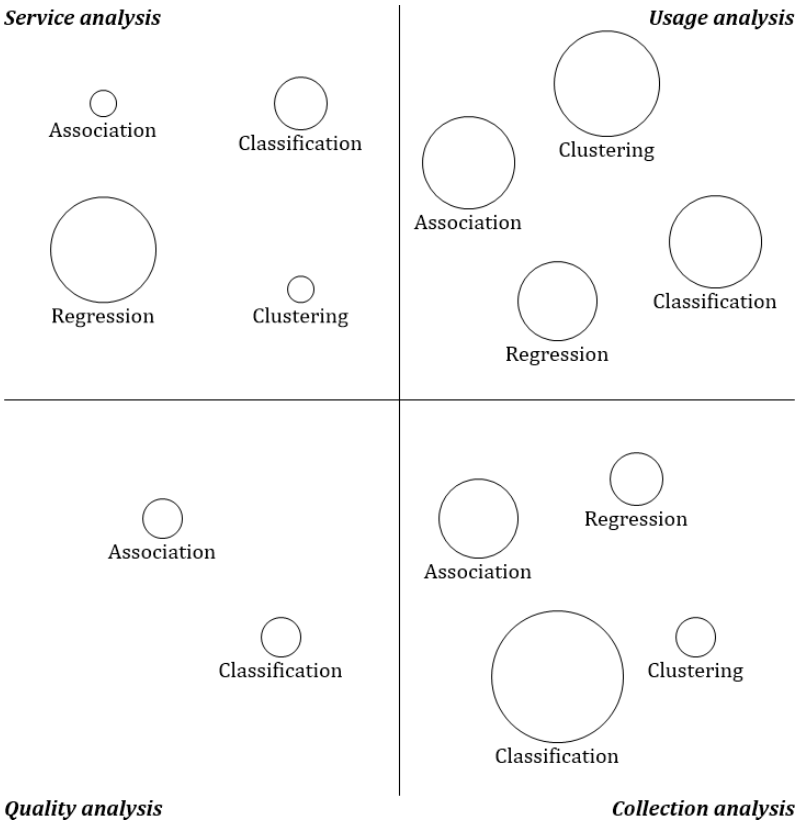

Figure 4: Classification of data mining applications based on the holistic evaluation matrix

Table 4 shows the distribution of articles by data mining techniques. Among 14 data mining techniques, which have been applied in libraries, logistic regression is the most commonly used technique (six out of 41 articles), followed closely by association rules, decision/classification tree, linear regression, and logical analysis of data (five articles each). Among the top 10 data mining techniques, log analysis is described in four articles, and statistical analysis, k-means clustering and pattern based clustering are each described in three articles.

Table 4

Distribution of articles by data mining techniques and library holistic quadrants

Data mining techniques	Service analysis	Quality analysis	Collection analysis	Usage analysis	Frequency	Number of articles	Percentage (%)
Logistic regression	4		2	2	*8*	*6*	*15*
Association rules	1		3	4	*8*	*5*	*11*
Decision/Classification tree			5	1	*6*	*5*	*11*
Logical analysis of data	3		2	2	*7*	*5*	*11*
Linear regression	4	1	2	4	*11*	*5*	*11*
Log analysis			1	4	*5*	*4*	*9*

K-means algorithm	1			3	*4*	*3*		*7*
Pattern based clustering			1	3	*4*	*3*		*7*
Statistical analysis			1	2	*3*	*3*		*7*
Hierarchical cluster analysis				2	*2*	*2*		*4*
Neutral network		2	1	1	*4*	*2*		*4*
Bibliometric analysis			1		*1*	*1*		*2*
Memory-based reasoning			1		*1*	*1*		*2*
Regression analysis	1	1			*2*	*1*		*2*
Count: 14	*14*	*4*	*20*	*28*		*Total:*	*41*	*100.00*

* Remark: Each article may have used more than one data mining technique and may have been implemented in more than one library holistic quadrant

Although logistic regression is implemented in the majority of articles, linear regression is the technique with the highest frequency among library quadrants, and in turn the unique implemented in all library holistic quadrants. Furthermore, regression models are the techniques that have been constantly implemented from the beginning. Collection and usage analyses both utilize 11 different types of algorithms (out of 14). Nevertheless, usage analysis is the quadrant that most data mining techniques have employed (28 techniques).

4.2 Distribution of articles by year of publication

The distribution of articles by year of publication and country of implementation is depicted in Table 5. One of the first reports on the use of data mining techniques in libraries is provided by Banerjee (1998). The author describes data mining functionalities, raises issues with their use, and discusses prerequisites for successfully utilizing data mining in libraries. Over the last six years (2009 to 2014), a positive increasing trend can be observed (see Figure 5), except for 2013 in which only two articles were published.

Table 5

Distribution of articles by year of publication and country of implementation

	Australia	China	Czech Republic	Germany	Greece	Netherlands	Taiwan	Thailand	UK	USA	*Total*
1998										1	*1*
1999											*0*
2000											*0*
2001											*0*
2002										2	*2*

2003					1		3			2	*6*
2004							1			1	*2*
2005											*0*
2006				2		1			1		*4*
2007	1									1	*2*
2008											*0*
2009										2	*2*
2010										3	*3*
2011								1		3	*4*
2012					2		2			2	*6*
2013		1								1	*2*
2014	1	1	1							4	*7*
Total	*2*	*2*	*1*	*2*	*3*	*1*	*6*	*1*	*1*	*22*	*41*

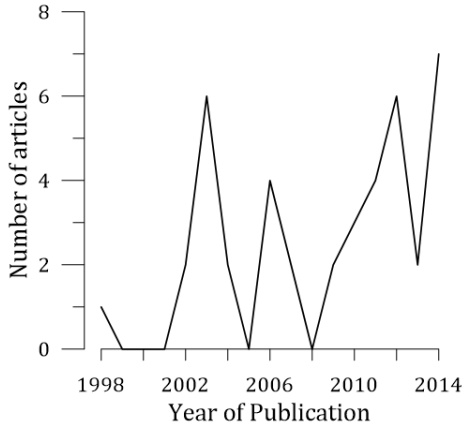

Figure 5: Evolution of number of articles per year

Among the 16-year publication analysis, it is remarkable that in 2003 and 2006, the number of publications increased from an average of about two papers per year to six and four papers respectively, when compared to other years. Despite the fact that the first publication on the topic was on 1998, this trend reflects that the true first efforts of implementation of data mining functions in libraries were carried on from 2002. In 2003, three out of six articles report a case study implementation in Taiwan and two articles in the USA. In Taiwan, Kao et al. (2003), Wu (2003), and Wu et al. (2004) lead the implementation of data mining techniques in libraries by developing a knowledge management framework that utilizes data mining of circulation data to assess use of materials by particular academic

departments in their subject areas. The techniques utilized in these studies are decision tree and association rules. In the USA, important to highlight is the study presented by Nicholson (2003a) that compares the effectiveness of four different data mining functions: logistic regression, memory-based reasoning, decision/classification tree and neural networks to discover Web-based scholarly research works. Moreover, in 2006, two out of four articles report a case study implementation in Germany, all implemented by Decker and colleagues (Decker & Hermelbracht, 2006; Decker & Höppner, 2006).

Data analysis and characterization are the most used techniques, except for the years 2002 – 2004, in which more formal data mining techniques are implemented. Quality is the only quadrant that has not employed these classical techniques as data mining tools. Important to note is that in the year 2014, articles have utilized the majority of data mining techniques, being 70% related to more formal procedures.

Usage analyses show an average of 1.4 publications per year throughout the period analyzed (see Figure 6). Strikingly, is also that in 2014, an increased interest in usage analyses is observed, since 6 out of 7 articles published in this year utilize several data mining functions, such as regression analysis, data characterization and cluster analysis, to analyze the usage of library e-resources, as well as to understand the information seeking behavior of users. Collection analyses are the second highest starting from 2003 onwards, especially in 2003 and 2012 with four out of seven articles and four out of eight articles published in those years respectively. Only isolated attempts to implement data mining functions in quality analyses can be observed in 2002, 2006, 2011, and 2012. Unfortunately, no studies have been reported on the use of data mining in quality analyses since 2012. Finally, articles focused on the use of data mining functions in service analyses emerged from 2006 onwards.

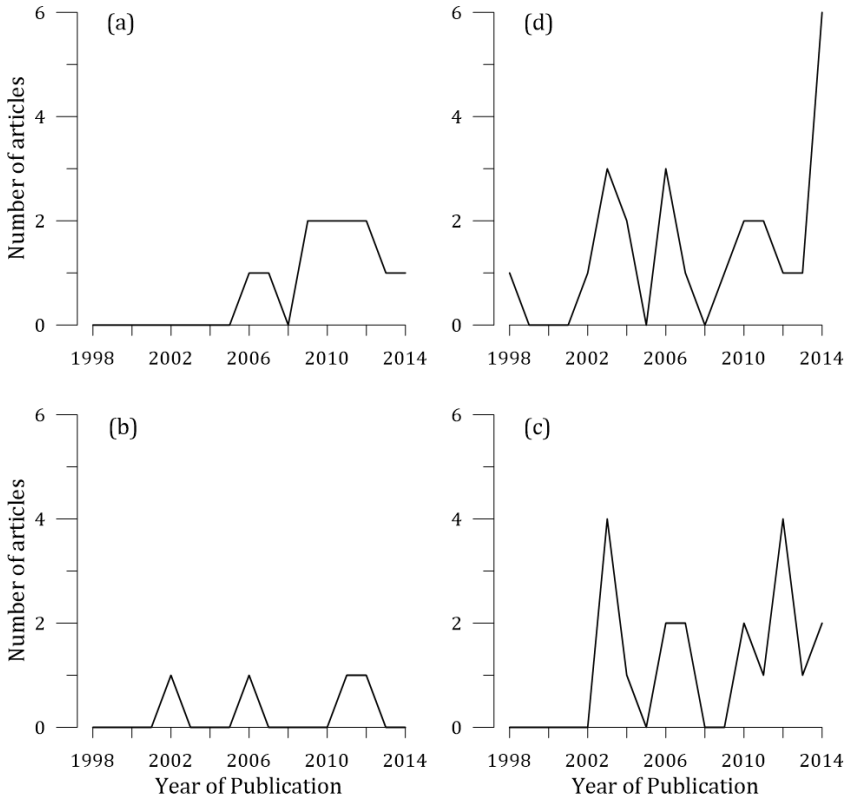

Figure 6: Chronological evolution of articles by library holistic evaluation. **(a)** Service analysis; **(b)** Quality analysis; **(c)** Collection analysis; **(d)** Usage analysis

4.3 Distribution of articles by country of implementation

Figure 7 shows the distribution of articles by country of implementation. Interesting is to highlight that the United States led by far among the list of countries that reported the implementation of data mining techniques (54% of case studies). In second, followed Taiwan, with a population of less than one tenth of the USA, with 15% of case studies implemented (six in total), and third Greece with 7% of case studies implemented (three out of 41). It is worth noting that four out of four articles covering both service and collection analyses are implemented in US academic libraries.

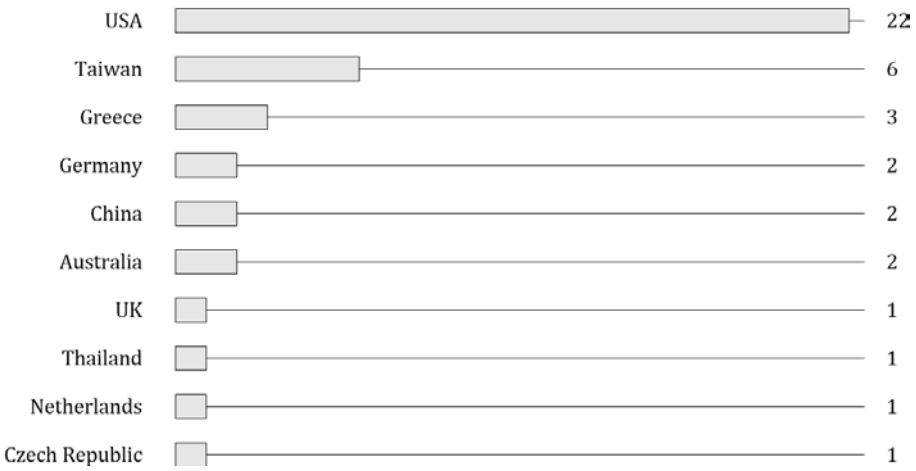

Figure 7: Distribution of articles by country of implementation

4.4 Distribution of articles by journal in which the articles were published

Figure 8 shows the top six journals, which contain the highest number of research articles. Articles related to application of data mining techniques in libraries are distributed across 27 journals. These findings indicate that scientific contributions in this research area scattered across a high range of journals (average of 1.52 articles per journal), particularly related to computer science, and information and library management. The top six journals which contain the highest number of research articles, contain almost 50% (20 out of 41 articles) of the total number of articles published.

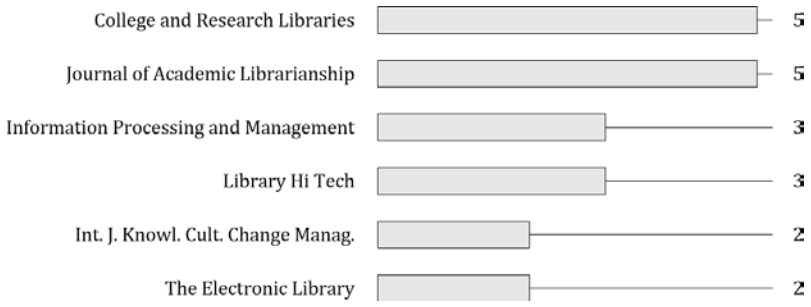

Figure 8: Distribution of articles in the top 6 journals

Of these, ‘‘College and Research Libraries” the official scholarly research journal of the Association of College & Research Libraries, and the “Journal of Academic Librarianship" both focused on problems and issues relevant to college and university libraries, each containing over 12% (five out of 41 articles)

of the total number of articles published, followed by the 'Information Processing and Management" and "Library Hi Tech" with three articles each, and the "International Journal of Knowledge, Culture and Change Management" and "The Electronic Library" journal with two articles each. All related to libraries except for the third and fifth ranked journal; the third journal is IT related, while the fifth is Management related.

The four databases (LISA, LISTA, WoS, and Scopus) were rechecked to determine where the articles were indexed. Scopus is the multidisciplinary database that provides the best coverage of journal articles identified in this study with 39 articles in total, followed by WoS with 34 articles found. LISA and LISTA index 26 and 27 articles out of 41 respectively. Evidently, the combination of the online databases allowed for the gathering the 41 analyzed articles.

4.5 Distribution of articles by library type analyzed

Figure 9 shows the distribution of articles by type of library analyzed: physical, digital, or both, through data mining techniques. Of the 41 articles, 41% (17 articles) are related to the use of data mining techniques in both digital and physical libraries, and 37% (15 articles) are related to the use of data mining techniques in digital libraries. This result is not unexpected, and confirms the natural transition and evolving trend of shifting the focus from physical to digital collection and services. Important to note is that not all the articles clearly specified the type of library analyzed, therefore, a certain subjectivity degree can be present. In addition, three articles reported by Yi (2009, 2011; 2012) are not specifically focused on a specific library type, since all examine how academic library directors plan and manage change in information technology and the factors influencing the planning and management approaches used.

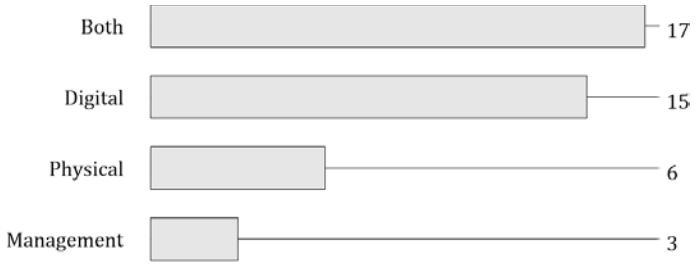

Figure 9: Distribution of articles by type of library analyzed

5. Limitations

The methodology that is employed in the literature review and classification of data mining techniques in libraries has some limitations. The first is that the study analyzes articles extracted based on specific keywords such as “data mining”, “case studies”, “academic librar*”, “university librar*” and “bibliomining”. Articles which mentioned the application of data mining techniques in academic libraries without these keywords may have been omitted during the retrieval process. The second is that findings are based on data collected only from academic journals, so other materials which may contain more case studies on data mining applications might have been excluded, and the third is the limited number of databases used (two multidisciplinary and two LIS oriented), of these, only the journals in the particular databases that were searched were included. However, although this limitation could mean that the review is not exhaustive, the authors believe that it is comprehensive by providing reasonable insights into the work being accomplished in the area, and also because the databases selected are the most important journal databases in their corresponding domain. The fourth possible limitation is that the articles’ classification process was subjective; nevertheless, research triangulation allowed for a reduction of this risk. Finally, the last limitation is that the study includes only English publications; by so doing, this restriction could jeopardize the analysis since surely more research regarding the application of data mining techniques in libraries is being discussed and published in other languages.

6. Conclusion and research implications

The application of data mining techniques in libraries is an emerging trend that has captured the attention of practitioners and academics in order to understand patterns of behavior of library users and staff, and patterns of information on resource usage throughout the library. The aim in this literature review has been to facilitate and ease the interested reader or practitioner's introduction to the use of data mining in libraries. To do so, the article presents a comprehensive literature review on the implementation of data mining techniques in libraries with a special focus on the case studies published over the period 1998-2014. Forty-one papers were identified, analyzed and classified along the four quadrants of the holistic evaluation matrix that analyze services, quality, collection, and usage behavior in libraries; and the main data mining functions, which are, clustering, association, classification, and regression. Although this literature review cannot claim to be exhaustive, it does highlight important implications, as well as insights into the state-of-the-art. For instance:

- Nicholson's idea of coining the term "bibliomining" to refer to the use of data mining in libraries (Nicholson & Stanton, 2003) was an important contribution in classifying this rapidly emerging topic. Data mining in libraries can be defined as the core of a larger process dubbed as bibliomining. Thus, the use of data mining to examine library data records might be aptly termed bibliomining. Unfortunately, in practice, only few researchers have used the "bibliomining" term in their publications (3 articles), and consequently, it cannot be considered as a standard word.
- According to past publication rates and the increasing interest in the use of data mining tools in libraries, practical research will increase significantly in this area in the future, and consequently, a significant increase in research and published literature is expected.
- Among the reviewed articles, all of them use one or two data mining techniques to analyze only one or two library holistic quadrants, and just one case study in the literature, by Emmons and Wilkinson (2011), has reported a case study covering the analysis of three library holistic quadrants: process, collection, and usage. None of the case studies cover the four library evaluation quadrants. Knowing that a combination of data mining models and library evaluation quadrants is often required to solve,

support, or forecast the effects of library strategies, library directors should include more data mining functions to support their holistic-based decision-making.

- The majority of the reviewed articles relate to usage analysis. Of these, about 38% (nine out of 24 articles) discuss data analysis techniques such as logical analysis of data and analysis of logs and statistics, 33% (eight articles) use cluster analysis, 29% (seven articles) utilize supervised learning tools and the remaining 17% (four articles) analyze dependences through association rules. The main library aspects covered through these studies are the interaction of library users with the system, the usability of library websites, and the users' categorization based on the usage interaction with the system and collection.
- Only a few articles of the 41 reviewed are related to quality analysis (four articles in total). The small number of research was somewhat surprising given that libraries have a long history on collecting statistics to answer users' queries, and thus monitor service quality (Horn & Owen, 2009); however, this topic is scarcely covered in the LIS literature since only a limited number of articles have reported the usage of sophisticated quality analysis as shown in this study. Further research needs to be conducted in this area, especially in quality control or in considering quality as an important factor when implementing data mining functions in other library aspects.
- Findings indicate that service analysis is slowly emerging as a possible new domain from this research. This library aspect is a crucial element for successful decision-making, especially due to increasingly difficult times, characterized by budget constraints and dynamic services. More than ever, libraries need to demonstrate that their processes and inputs such as facilities, expenditures, and staffing are considered relevant and worthwhile in their outputs through data on services and people served. Therefore, more research is highly recommended on the use of data mining techniques in the analysis of service performance in both digital and physical environments.
- During the research a common theme that has emerged was the appropriated definition of "data mining". Actually, the concept is difficult to explain and several authors opine that the term is a

misnomer and a buzzword (Han et al., 2011; J. Wu, 2012). In this study, 10 out of 11 articles (almost 25% of 41 articles) implementing data analysis and characterization approaches, such as logical analysis of data, and analyses of statistics, logs and bibliometrics, include as part of their topic terms (title, keywords or even descriptors) the words “data mining”. The reasoning behind the inclusion of these articles, which can be argued to be not data mining techniques, is to highlight the overuse of the data mining concept in the LIS literature. Therefore, to benefit from the advantages of data mining, it is recommended that further studies be conducted utilizing more enhanced techniques that have not been documented previously such as super vector machines and ensemble methods.

- Among the 41 case studies reviewed for this article, 14 articles utilize classification models and 11 use regression techniques to assist in library decision-making. Laggards are the implementation of unsupervised algorithms such as association and clustering models (eight out of 41 articles each). Knowing that unsupervised learning allows finding hidden structure in unlabeled data, as well as allows spotting salient correlations and connections between data points that are not evident for humans, the implementation of further association and clustering algorithms is highly recommended.
- Association rules and decision/classification trees rank after logistic regression in popularity of application in libraries. The logic of both techniques can be followed more easily by librarians and information specialists. Therefore, the two techniques are highly recommended for non-experts in data mining techniques.
- The top six journals which contain almost 50% of the total number of articles published on the application of data mining techniques in academic libraries are: College and Research Libraries, Journal of Academic Librarianship, Information Processing and Management, Library Hi Tech, International Journal of Knowledge, Culture and Change Management, and The Electronic Library. Scopus is the multidisciplinary database that indexes almost all articles identified in this study.
- The majority of research articles have been implemented in the United States.

- Findings indicate that the attention of implementing data mining techniques in library management literature has mainly been directed towards digital collection and e-services (37%), and less towards physical collection (15%). This is not surprising as digital libraries are becoming more and more prevalent worldwide.

Acknowledges

This research project was funded by the Flemish Interuniversity Council (VLIR-IUC), the National Secretariat of Higher Education, Science, Technology and Innovation of Ecuador (SENESCYT); and supported by the CEPRA VII project "Plataforma de integración, publicacíon y consulta integrada de recursos bibliogŕaficos en la Web Semantica" funded by the Ecuadorian Consortium for Advanced Internet Development (CEDIA). The authors thank Andres Auquilla for the fruitful discussions on data mining techniques trends, and Paul Vanegas for reviewing some drafts of this article.

References

ACRL Research Planning and Review Committee. (2014). Top trends in academic libraries: A review of the trends and issues affecting academic libraries in higher education. *College & Research Libraries News*, *75*(6), 294–302.

Ahmad, P., Brogan, M., & Johnstone, M. N. (2014). The e-book power user in academic and research libraries: Deep log analysis and user customisation. *Australian Academic & Research Libraries*, *45*(1), 35–47. http://doi.org/10.1080/00048623.2014.885374

Association of College and Research Libraries. (2010). *Value of academic libraries: A comprehensive research review and report*. Chicago, USA: Association of College and Research Libraries.

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries*, *18*(10), 28–31.

Blecic, D. D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H., & Weller, A. C. (1998). Using transaction log analysis to improve OPAC retrieval results. *College & Research Libraries*, *59*(1), 39–50. http://doi.org/10.5860/crl.59.1.39

Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, *8*(6). http://doi.org/10.1045/june2002-bollen

Bracke, P. J. (2004). Web usage mining at an academic health sciences library: an exploratory study. *Journal of the Medical Library Association*, *92*(4), 421–428.

Chang, C.-C., & Chen, R.-S. (2006). Using data mining technology to solve classification problems: A case study of campus digital library. *Electronic Library, The*, *24*(3), 307–321. http://doi.org/10.1108/02640470610671178

Chen, S. Y., & Liu, X. (2004). The contribution of data mining to information science. *Journal of Information Science*, *30*(6), 550–558. http://doi.org/10.1177/0165551504047928

Decker, R., & Hermelbracht, A. (2006). Planning and evaluation of new academic library services by means of Web-based conjoint analysis. *The Journal of Academic Librarianship*, *32*(6), 558–572. http://doi.org/10.1016/j.acalib.2006.06.016

Decker, R., & Höppner, M. (2006). Strategic planning and customer intelligence in academic libraries. *Library Hi Tech*, *24*(4), 504–514. http://doi.org/10.1108/07378830610715374

Emmons, M., & Wilkinson, F. C. (2011). The Academic Library Impact on Student Persistence. *College & Research Libraries*, crl–74r1.

Fagan, J. C. (2014). The Effects of Reference, Instruction, Database Searches, and Ongoing Expenditures on Full-text Article Requests: An Exploratory Analysis. *The Journal of Academic Librarianship*, *40*(3–4), 264–274. http://doi.org/10.1016/j.acalib.2014.04.002

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM*, *39*(11), 27–34. http://doi.org/10.1145/240455.240464

Finnell, J., & Fontane, W. (2010). Reference Question Data Mining: A Systematic Approach to Library Outreach. *Reference & User Services Quarterly*, *49*(3), 278–286.

Girija, N., & Srivatsa, S. K. (2006). A research study: Using Data Mining in knowledge base business strategies. *Information Technology Journal*, *5*(3), 590–600. http://doi.org/10.3923/itj.2006.590.600

Gonzalez, R., Llopis, J., & Gasco, J. (2013). Information systems offshore outsourcing: managerial conclusions from academic research. *International Entrepreneurship and Management Journal*, *9*(2), 229–259. http://doi.org/10.1007/s11365-013-0250-y

Hájek, P., & Stejskal, J. (2014). Library user behavior analysis - Use in economics and management. *Wseas Transactions on Business and Economics*, *11*, 107–116.

Hand, D. J. (1998). Data Mining: Statistics and more? *The American Statistician*, *52*(2), 112–118. http://doi.org/10.1080/00031305.1998.10480549

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and techniques* (3rd ed.). Elsevier.

Horn, A., & Owen, S. (2009). Mind the gap 2014 : research to inform the next five years of library development. In *Innovate, collaborate : conference proceedings EDUCAUSE Australasia 2009* (pp. 1–12). Perth, Western Australia. Retrieved from http://dro.deakin.edu.au/view/DU:30016487

Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information & Management*, *38*(1), 1–13. http://doi.org/10.1016/S0378-7206(00)00051-3

Kao, S.-C., Chang, H.-C., & Lin, C.-H. (2003). Decision support for the academic library acquisition budget allocation via circulation database mining. *Information Processing & Management*, *39*(1), 133–147. http://doi.org/10.1016/S0306-4573(02)00019-5

King, J. D., Li, Y., Tao, X., & Nayak, R. (2007). Mining world knowledge for analysis of search engine content. *Web Intelligence and Agent Systems*, *5*(3), 233–253.

Kononenko, I., & Kukar, M. (2007). *Machine Learning and Data Mining*. Elsevier.

Koulouris, A., & Kapidakis, S. (2012). Policy route map for academic libraries' digital content. *Journal of Librarianship and Information Science*, *44*(3), 163–173. http://doi.org/10.1177/0961000612444299

Leonard, M. F., Haas, S. C., & Kisling, V. N. (2010). Metrics and science monograph collections at the marston science library, University of Florida. *Issues in Science and Technology Librarianship*, *62*. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=3706724

Li, X. (2014). An Algorithm for Mining Frequent Itemsets from Library Big Data. *Journal of Software*, *9*(9), 2361–2365. http://doi.org/10.4304/jsw.9.9.2361-2365

Lunfeng, G., Huan, L., & Li, Z. (2012). The application of association rules of data mining in book-lending service. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 761–764). http://doi.org/10.1109/FSKD.2012.6233921

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2, Part 2), 2592–2602. http://doi.org/10.1016/j.eswa.2008.02.021

Nicholas, D., Huntington, P., Monopoli, M., & Watkinson, A. (2006). Engaging with scholarly digital libraries (publisher platforms): The extent to which "added-value" functions are used. *Information Processing & Management*, *42*(3), 826–842. http://doi.org/10.1016/j.ipm.2005.03.019

Nicholson, S. (2003a). Bibliomining for automated collection development in a digital library setting: Using data mining to discover Web-based scholarly research works. *Journal of the American Society for Information Science and Technology*, *54*(12), 1081–1090. http://doi.org/10.1002/asi.10313

Nicholson, S. (2003b). The bibliomining process: Data Warehousing and Data Mining for library decision-making. *Information Technology and Libraries*, *22*(4), 146–151.

Nicholson, S. (2004). A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Journal of Documentation*, *60*(2), 164–182. http://doi.org/10.1108/00220410410522043

Nicholson, S. (2006a). Approaching librarianship from the data: Using bibliomining for evidence-based librarianship. *Library Hi Tech*, *24*(3), 369–375. http://doi.org/10.1108/07378830610692136

Nicholson, S. (2006b). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing & Management*, *42*(3), 785–804. http://doi.org/10.1016/j.ipm.2005.05.008

Nicholson, S., & Stanton, J. (2006). Bibliomining for library decision-making. In *Encyclopedia of Data Warehousing and Mining* (Second Edition, pp. 100–105). Retrieved from http://www.igi-global.com/chapter/encyclopedia-data-warehousing-mining/10591

Nicholson, S., & Stanton, J. M. (2003). Gaining strategic advantage through Bibliomining: Data Mining for management decisions in corporate, special, digital, and traditional libraries. In *Organizational data mining: Leveraging enterprise data resources for optimal performanc*. Hershey, PA: Idea Group Publishing. Retrieved from http://arizona.openrepository.com/arizona/bitstream/10150/106383/1/Nicholson_3.pdf

Papatheodorou, C., Kapidakis, S., Sfakakis, M., & Vassiliou, A. (2003). Mining user communities in digital libraries. *Information Technology and Libraries*, *22*(4), 152–157.

Papavlasopoulos, S., & Poulos, M. (2012). Neural network design and evaluation for classifying library indicators using personal opinion of expert. *Library Management*, *33*(4/5), 261–271. http://doi.org/10.1108/01435121211242308

Prakash, K., Chand, P., & Gohel, U. (2004). Application of Data Mining in library and information services (pp. 168–177). Presented at the 2nd Convention PLANNER, Manipur Uni., Imphal: INFLIBNET Centre, Ahmedabad. Retrieved from http://shodhganga.inflibnet.ac.in/dxml/handle/1944/435

Pu, H., & Yang, C. (2003). Enriching user-oriented class associations for library classification schemes. *The Electronic Library*, *21*(2), 130–141. http://doi.org/10.1108/02640470310470507

Robin R. Sewell. (2013). Who is following us? Data mining a library's Twitter followers. *Library Hi Tech*, *31*(1), 160–170. http://doi.org/10.1108/07378831311303994

Samson, S. (2014). Usage of e-resources: Virtual value of demographics. *The Journal of Academic Librarianship*, *40*(6), 620–625. http://doi.org/10.1016/j.acalib.2014.10.005

Shieh, J.-C. (2010). The integration system for librarians' bibliomining. *Electronic Library, The*, *28*(5), 709–721. http://doi.org/10.1108/02640471011081988

Shieh, J.-C. (2012). From website log to findability. *Electronic Library, The*, *30*(5), 707–720. http://doi.org/10.1108/02640471211275747

Shreeves, S. L., Kaczmarek, J. S., & Cole, T. W. (2003). Harvesting cultural heritage metadata using the OAI Protocol. *Library Hi Tech*, *21*(2), 159–169. http://doi.org/10.1108/07378830310479802

Siguenza-Guzman, L., Van Den Abbeele, A., Vandewalle, J., Verhaaren, H., & Cattrysse, D. (2015). A holistic approach to supporting academic libraries in resource allocation processes. *The Library Quarterly: Information, Community, Policy (in Press)*.

Siriprasoetsin, P., Tuamsuk, K., & Vongprasert, C. (2011). Factors affecting customer relationship management practices in Thai academic libraries. *The International Information & Library Review*, *43*(4), 221–229. http://doi.org/10.1016/j.iilr.2011.10.008

Soria, K. M., Fransen, J., & Nackerud, S. (2014). Stacks, Serials, Search Engines, and Students' Success: First-Year Undergraduate Students' Library Use, Academic Achievement, and Retention. *The Journal of Academic Librarianship*, *40*(1), 84–91. http://doi.org/10.1016/j.acalib.2013.12.002

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its applications*. Springer.

Tempelman-Kluit, N., & Pearce, A. (2014). Invoking the User from Data to Design. *College & Research Libraries*, *75*(5), 616–640. http://doi.org/10.5860/crl.75.5.616

Todorinova, L., Huse, A., Lewis, B., & Torrence, M. (2011). Making Decisions: Using Electronic Data Collection to Re-Envision Reference Services at the USF Tampa Libraries. *Public Services Quarterly*, *7*(1-2), 34–48. http://doi.org/10.1080/15228959.2011.572780

Tosaka, Y., & Weng, C. (2011). Reexamining Content-Enriched Access: Its Effect on Usage and Discovery. *College & Research Libraries*, *72*(5), 412–427. http://doi.org/10.5860/crl-137

Walters, W. H. (2007). A Regression-based Approach to Library Fund Allocation. *Library Resources & Technical Services*, *51*(4), 263–278.

Weiner, S. (2009). The Contribution of the Library to the Reputation of a University. *The Journal of Academic Librarianship*, *35*(1), 3–13. http://doi.org/10.1016/j.acalib.2008.10.003

Whitmire, E. (2002). Academic library performance measures and undergraduates' library use and educational outcomes. *Library & Information Science Research*, *24*(2), 107–128. http://doi.org/10.1016/S0740-8188(02)00108-1

Will, N. (2006). Data-mining: Improvement of university library services. *Technological Forecasting and Social Change*, *73*(8), 1045–1050. http://doi.org/10.1016/j.techfore.2006.05.006

Wright, S., & White, L. S. (2007). Library Assessment: SPEC Kit 303. *Association of Research Libraries*, 14.

Wu, C.-H. (2003). Data mining applied to material acquisition budget allocation for libraries: design and development. *Expert Systems with Applications*, *25*(3), 401–411. http://doi.org/10.1016/S0957-4174(03)00065-4

Wu, C.-H., Lee, T.-Z., & Kao, S.-C. (2004). Knowledge discovery applied to material acquisitions for libraries. *Information Processing & Management*, *40*(4), 709–725. http://doi.org/10.1016/j.ipm.2003.08.010

Wu, J. (2012). *Advances in K-means Clustering: A Data Mining Thinking*. Springer Science & Business Media.

Yang, S.-T. (2012). An active recommendation approach to improve book-acquisition process. *International Journal of Electronic Business Management*, *10*(2), 163–73.

Yi, Z. (2009). The management of change in information technology: Approaches of academic library directors in the United States. *International Journal of Knowledge, Culture and Change Management*, *9*(11), 109–130.

Yi, Z. (2011). Planning change in the information age: Approaches of academic library directors in the United States. *International Journal of Knowledge, Culture and Change Management*, *10*(12), 155–176.

Yi, Z. (2012). Conducting meetings in the change process: Approaches of academic library directors in the United States. *Library Management*, *33*(1/2), 22–35. http://doi.org/10.1108/01435121211203293

Zhang, Q. S., & Wang, X. Y. (2013). Research of Personalized Information Service Based on Association Rules. *Advanced Materials Research*, *760-762*, 1800–1803. http://doi.org/10.4028/www.scientific.net/AMR.760-762.1800

Zweibel, S., & Lane, Z. B. (2012). Probing the effects of policy changes by evaluating circulation activity data at Columbia University Libraries. *The Serials Librarian*, *63*(1), 17–27. http://doi.org/10.1080/0361526X.2012.687850

Appendix A

References of the empirical studies implementing data mining techniques in libraries

Ahmad, P., Brogan, M., & Johnstone, M. N. (2014). The e-book power user in academic and research libraries: Deep log analysis and user customisation. *Australian Academic & Research Libraries*, 45(1), 35–47. http://doi.org/10.1080/00048623.2014.885374

Blecic, D. D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H., & Weller, A. C. (1998). Using transaction log analysis to improve OPAC retrieval results. *College & Research Libraries*, 59(1), 39–50. http://doi.org/10.5860/crl.59.1.39

Bollen, J., & Luce, R. (2002). Evaluation of digital library impact and user communities by analysis of usage patterns. *D-Lib Magazine*, 8(6). http://doi.org/10.1045/june2002-bollen

Bracke, P. J. (2004). Web usage mining at an academic health sciences library: an exploratory study. *Journal of the Medical Library Association*, 92(4), 421–428.

Decker, R., & Hermelbracht, A. (2006). Planning and evaluation of new academic library services by means of Web-based conjoint analysis. *The Journal of Academic Librarianship*, 32(6), 558–572. http://doi.org/10.1016/j.acalib.2006.06.016

Decker, R., & Höppner, M. (2006). Strategic planning and customer intelligence in academic libraries. *Library Hi Tech*, 24(4), 504–514. http://doi.org/10.1108/07378830610715374

Emmons, M., & Wilkinson, F. C. (2011). The Academic Library Impact on Student Persistence. *College & Research Libraries*, crl–74r1.

Fagan, J. C. (2014). The Effects of Reference, Instruction, Database Searches, and Ongoing Expenditures on Full-text Article Requests: An Exploratory Analysis. *The Journal of Academic Librarianship*, 40(3–4), 264–274. http://doi.org/10.1016/j.acalib.2014.04.002

Finnell, J., & Fontane, W. (2010). Reference Question Data Mining: A Systematic Approach to Library Outreach. *Reference & User Services Quarterly*, 49(3), 278–286.

Hájek, P., & Stejskal, J. (2014). Library user behavior analysis - Use in economics and management. *Wseas Transactions on Business and Economics*, 11, 107–116.

Kao, S.-C., Chang, H.-C., & Lin, C.-H. (2003). Decision support for the academic library acquisition budget allocation via circulation database mining. *Information Processing & Management*, 39(1), 133–147. http://doi.org/10.1016/S0306-4573(02)00019-5

King, J. D., Li, Y., Tao, X., & Nayak, R. (2007). Mining world knowledge for analysis of search engine content. *Web Intelligence and Agent Systems*, 5(3), 233–253.

Koulouris, A., & Kapidakis, S. (2012). Policy route map for academic libraries' digital content. *Journal of Librarianship and Information Science*, 44(3), 163–173. http://doi.org/10.1177/0961000612444299

Leonard, M. F., Haas, S. C., & Kisling, V. N. (2010). Metrics and science monograph collections at the marston science library, University of Florida. *Issues in Science and Technology Librarianship*, 62. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=3706724

Li, X. (2014). An Algorithm for Mining Frequent Itemsets from Library Big Data. *Journal of Software*, 9(9), 2361–2365. http://doi.org/10.4304/jsw.9.9.2361-2365

Nicholas, D., Huntington, P., Monopoli, M., & Watkinson, A. (2006). Engaging with scholarly digital libraries (publisher platforms): The extent to which "added-value" functions are used. *Information Processing & Management*, 42(3), 826–842. http://doi.org/10.1016/j.ipm.2005.03.019

Nicholson, S. (2003). Bibliomining for automated collection development in a digital library setting: Using data mining to discover Web *Journal of Library Research* works. *Society for Information Science and Technology*, 54(12), 1081–1090. http://doi.org/10.1002/asi.10313

Papatheodorou, C., Kapidakis, S., Sfakakis, M., & Vassiliou, A. (2003). Mining user communities in digital libraries. *Information Technology and Libraries*, 22(4), 152–157.

Papavlasopoulos, S., & Poulos, M. (2012). Neural network design and evaluation for classifying library indicators using personal opinion of expert. *Library Management*, 33(4/5), 261–271. http://doi.org/10.1108/01435121211242308

Pu, H., & Yang, C. (2003). Enriching user -oriented class schemes. *The Electronic Library*, 21(2), 130–141. http://doi.org/10.1108/02640470310470507

Robin R. Sewell. (2013). Who is following us? Data mining a library's Twitter followers. *Library Hi Tech*, 31(1), 160–170. http://doi.org/10.1108/07378831311303994

Samson, S. (2014). Usage of e-resources: Virtual value of demographics. *The Journal of Academic Librarianship*, 40(6), 620–625. http://doi.org/10.1016/j.acalib.2014.10.005

Shieh, J.-C. (2012). From website log to findability. *Electronic Library, The*, 30(5), 707–720. http://doi.org/10.1108/02640471211275747

Shreeves, S. L., Kaczmarek, J. S., & Cole, T. W. (2003). Harvesting cultural heritage metadata using the OAI Protocol. *Library Hi Tech*, 21(2), 159–169. http://doi.org/10.1108/07378830310479802

Siriprasoetsin, P., Tuamsuk, K., & Vongprasert, C. (2011). Factors affecting customer relationship management practices in Thai academic libraries. *The International Information & Library Review*, 43(4), 221–229. http://doi.org/10.1016/j.iilr.2011.10.008

Soria, K. M., Fransen, J., & Nackerud, S. (2014). Stacks, Serials, Search Engines, and Students' Success: First-Year Undergraduate Students' Library Use, Academic Achievement, and Retention. *The Journal of Academic Librarianship*, 40(1), 84–91. http://doi.org/10.1016/j.acalib.2013.12.002

Tempelman-Kluit, N., & Pearce, A. (2014). Invoking the User from Data to Design. *College & Research Libraries*, 75(5), 616–640. http://doi.org/10.5860/crl.75.5.616

Todorinova, L., Huse, A., Lewis, B., & Torrence, M. (2011). Making Decisions: Using Electronic Data Collection to Re-Envision Reference Services at the USF Tampa Libraries. *Public Services Quarterly*, 7(1-2), 34–48. http://doi.org/10.1080/15228959.2011.572780

Tosaka, Y., & Weng, C. (2011). Reexamining Content-Enriched Access: Its Effect on Usage and Discovery. *College & Research Libraries*, 72(5), 412–427. http://doi.org/10.5860/crl-137

Walters, W. H. (2007). A Regression-based Approach to Library Fund Allocation. *Library Resources & Technical Services*, 51(4), 263–278.

Weiner, S. (2009). The Contribution of the Library to the Reputation of a University. *The Journal of Academic Librarianship*, 35(1), 3–13. http://doi.org/10.1016/j.acalib.2008.10.003

Whitmire, E. (2002). Academic library performance measures and undergraduates' library use and educational outcomes. *Library & Information Science Research*, 24(2), 107–128. http://doi.org/10.1016/S0740-8188(02)00108-1

Will, N. (2006). Data-mining: Improvement of university library services. *Technological Forecasting and Social Change*, 73(8), 1045–1050. http://doi.org/10.1016/j.techfore.2006.05.006

Wu, C.-H. (2003). Data mining applied to material acquisition budget allocation for libraries: design and development. *Expert Systems with Applications*, 25(3), 401–411. http://doi.org/10.1016/S0957-4174(03)00065-4

Wu, C.-H., Lee, T.-Z., & Kao, S.-C. (2004). Knowledge discovery applied to material acquisitions for libraries. *Information Processing & Management*, 40(4), 709–725. http://doi.org/10.1016/j.ipm.2003.08.010

Yang, S.-T. (2012). An active recommendation approach to improve book-acquisition process. *International Journal of Electronic Business Management*, 10(2), 163–73.

Yi, Z. (2009). The management of change in information technology: Approaches of academic library directors in the United States. *International Journal of Knowledge, Culture and Change Management*, 9(11), 109–130.

Yi, Z. (2011). Planning change in the information age: Approaches of academic library directors in the United States. *International Journal of Knowledge, Culture and Change Management*, 10(12), 155–176.

Yi, Z. (2012). Conducting meetings in the change process: Approaches of academic library directors in the United States. *Library Management*, 33(1/2), 22–35. http://doi.org/10.1108/01435121211203293

Zhang, Q. S., & Wang, X. Y. (2013). Research of Personalized Information Service Based on Association Rules. *Advanced Materials Research*, 760-762, 1800–1803. http://doi.org/10.4028/www.scientific.net/AMR.760-762.1800

Zweibel, S., & Lane, Z. B. (2012). Probing the effects of policy changes by evaluating circulation activity data at Columbia University Libraries. *The Serials Librarian*, 63(1), 17–27. http://doi.org/10.1080/0361526X.2012.687850